Measuring the Reliability of Causal Probing Methods: Tradeoffs, Limitations, and the Plight of Nullifying Interventions

Marc E. Canby* **Adam Davies*** **Chirag Rastogi** **Julia Hockenmaier**
Siebel School of Computing and Data Science
The Grainger College of Engineering
University of Illinois Urbana-Champaign
{marcec2, adavies4,chiragr2,juliahmr}@illinois.edu

Abstract

Causal probing aims to analyze foundation models by examining how intervening on their representation of various latent properties impacts their outputs. Recent works have cast doubt on the theoretical basis of several leading causal probing methods, but it has been unclear how to systematically evaluate the effectiveness of these methods in practice. To address this, we formally define and quantify two key causal probing desiderata: *completeness* (how thoroughly the representation of the target property has been transformed) and *selectivity* (how little non-targeted properties have been impacted). We introduce an empirical analysis framework to measure and evaluate these quantities, allowing us to make the first direct comparisons of the reliability of different families of causal probing methods (e.g., linear vs. nonlinear or counterfactual vs. nullifying interventions). We find that: (1) there is an inherent tradeoff between completeness and selectivity; (2) no leading probing method is able to consistently satisfy both criteria at once; (3) methods with more favorable tradeoffs have a more consistent impact on LLM behavior; and (4) nullifying interventions are far less complete than counterfactual interventions, suggesting that nullifying methods may not be an effective approach to causal probing.

1 Introduction

What latent properties do large language models (LLMs) learn to represent, and how do they leverage such representations? Causal probing aims to answer this question by intervening on a model's embedding representations of some property of interest (e.g., parts-of-speech), feeding the altered embeddings back into the LLM, and assessing how the model's behavior on downstream tasks changes [11, 27, 10, 30, 18, 7, 32]. However, it is only possible to draw nontrivial conclusions about the model's use of the latent property if we are confident that interventions have fully and precisely carried out the intended transformation [8]. Indeed, prior works have raised serious doubts about causal probing, finding that many intervention methods may have a large unintended impact on non-targeted properties [17], and that the original value of the property may still be recoverable [10, 24]. So far, it has been unclear how these doubts generalize to other types of interventions or how serious they are in practice, as there is no generally accepted approach for evaluating or comparing different methods.

Thus, our main goal in this work is to work toward a systematic understanding of the effectiveness and limitations of current causal probing methodologies. Specifically, we propose an empirical analysis framework to evaluate the *reliability* of causal probing according to two key desiderata:

1. *Completeness*: interventions should fully remove the representation of targeted properties.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

*These authors contributed equally to this work.

2. *Selectivity*: interventions should not impact non-targeted properties.

We define completeness and selectivity using "oracle probes" that enable measuring the impact of an intervention on both targeted and non-targeted properties. We apply our framework to several intervention methods, finding that they each show a clear tradeoff between these criteria, and that no method is able to satisfy them both simultaneously. We also show that the most complete and reliable interventions lead to the most consistent and substantial changes in LLM task performance. Finally, across all methods we study, we observe that counterfactual interventions are universally more complete and reliable than nullifying interventions. While this finding is consistent with earlier work criticizing nullifying interventions [10, 24, 17, 23], it has not previously been possible to directly compare them with counterfactual interventions, and evidence against nullifying interventions has been used to argue against causal probing more broadly [17]. However, our results show that the most serious limitations are generally only present for nullifying methods, and that counterfactual methods constitute a more reliable approach to causal probing.

2 Background and Related Work

Structural Probing The goal of *structural probing* [14, 21, 4] is to analyze which properties (e.g., part-of-speech, sentiment labels, etc.) are represented by a deep learning model (e.g., LLM) by training classifiers to predict these properties from latent embeddings [3]. Given, say, an LLM M, input token sequence $\mathbf{x} = (x_1, ..., x_N)$, and embeddings $\mathbf{h}^l = M_l(\mathbf{x})$ of input $\mathbf{x}$ at layer l of M, suppose Z is a latent property of interest that takes a discrete value $Z = z$ for input $\mathbf{x}$. The goal of structural probing is to train a classifier g_Z: $M_l(\mathbf{x}) \mapsto z$ to predict the value of Z from $\mathbf{h}^l$. On the most straightforward interpretation, if g_Z achieves high accuracy on the probe task, then the model is said to be "representing" Z [3]. An important criticism of such methodologies is that *correlation does not imply causation* – i.e., that simply because a given property can be predicted from embedding representations does not mean that the model is using the property in any way [13, 10, 3, 7].

Causal Probing A prominent response to this concern has been *causal probing*, which uses structural probes to remove or alter that property in the model's representation, and measuring the impact of such interventions on the model's predictions [10, 30, 18, 7]. Specifically, causal probing performs interventions $\mathrm{do}(Z)$ that modify M's representation of Z in the embeddings $\mathbf{h}^l$, producing $\hat{\mathbf{h}}^l$, where interventions can either encode a counterfactual value $Z = z'$ (denoted $\mathrm{do}(Z = z')$ where $z \neq z'$), or remove the representation of Z entirely (denoted $\mathrm{do}(Z = 0)$). Following the intervention, modified embeddings $\hat{\mathbf{h}}^l$ are fed back into M beginning at layer $l + 1$ to complete the forward pass, yielding intervened predictions $\Pr_M(\cdot|\mathbf{x}, \mathrm{do}(Z))$. Comparison with the original predictions $\Pr_M(\cdot|\mathbf{x})$ allows one to measure the extent to which M uses its representation of Z in computing them.

Causal Probing: Limitations Prior works have indicated that information about the target property that should have been completely removed may still be recoverable by the model [10, 24, 23], in which case interventions are not complete; or that most of the impact of interventions may actually be the result of collateral damage to correlated, non-targeted properties [17], in which case interventions are not selective. How seriously should we take such critiques? We observe several important shortcomings in each of these prior studies on the limitations of causal probing interventions:

1. These limitations have only been empirically demonstrated for the task of removing information about a target property from embeddings such that the model *cannot be fine-tuned to use the property for downstream tasks* [17, 24, 23]. But considering that the goal of causal probing is to interpret the behavior of an existing pre-trained model, the question is not whether models *can* be fine-tuned to use the property; it is whether models *already* use the property without task-specific fine-tuning, which has not been addressed in prior work. Do we observe the same limitations in this context?
2. These limitations have only been studied for linear nullifying interventions (e.g., [27, 26]), despite the recent proliferation of other causal probing methodologies, including nonlinear [30, 24, 29, 7] and counterfactual interventions [25, 30, 7] (see Section 4). Do we observe the same limitations for, e.g., nonlinear counterfactual interventions?

In this work, we answer both questions by providing a precise, quantifiable, and sufficiently general definition of completeness and selectivity that it is applicable to *all* such causal probing interventions, and carry out extensive experiments to evaluate representative methods from each category of interventions when applied to a pre-trained LLM as it performs a zero-shot prompt task. Note that,

while we are the first to define and measure the completeness and selectivity of causal probing interventions, Huang et al. [15] performs a broadly analogous analysis with respect to *interchange interventions*. Such interventions operate at the level of individual entities (e.g., France, Asia, or Europe) rather than general latent properties (e.g., part-of-speech) that have categorical values that are each taken by many different inputs, as studied in causal probing.

3 Evaluating Causal Probing Reliability

Recall that our main goal in this work is to evaluate intervention reliability in terms of completeness (completely transforming M's representation of some target property Z_i) and selectivity (minimally impacting M's representation of other properties $Z_j \neq Z_i$).[1] Given that we cannot directly inspect what value M encodes for any given property Z_i, it is necessary to introduce the notion of *oracle probes*, which we use to measure the extent to which interventions have fulfilled either criterion.

Oracle Probes We define an oracle probe o as a structural probe that returns a distribution $\Pr_o(Z|\mathbf{h})$ over the values of property Z, and we interpret $\Pr_o(Z = z|\mathbf{h})$ as the degree to which the model's embedding representations[2] $\mathbf{h}$ given natural-language input $\mathbf{x}$ encodes a belief that $\mathbf{x}$ has the property $Z = z$. So, if $\mathbf{h}$ encodes value $Z = \hat{z}$ with complete certainty, o should return a degenerate distribution $\Pr_o(Z|\mathbf{h}) = \mathbb{1}(Z = \hat{z})$, whereas we would expect a uniform distribution $\Pr_o(Z|\mathbf{h}) = \mathcal{U}(Z)$ if $\mathbf{h}$ does not encode property Z at all.[3] Naturally, a perfect oracle does not exist in practice, so any implementation must approximate it (see Section 4). However, a sufficiently high-quality oracle probe approximation enables us to estimate how well various intervention methods perform the desired modification.

Completeness If a counterfactual intervention $\mathrm{do}(Z = z')$ is perfectly *complete*, then it would produce a perfectly-intervened $\mathbf{h}^*_{Z=z'}$ that fully transforms $\mathbf{h}$ from encoding value $Z = z$ to encoding counterfactual value $Z = z' \neq z$. Thus, after performing the intervention, oracle o should emit $\Pr_o(Z = z'|\mathbf{h}^*_{Z=z'}) = P^*_Z(Z = z') = 1$. For nullifying interventions $\mathrm{do}(Z = 0)$, a perfectly complete representation $\mathbf{h}^*_{Z=0}$ should not encode Z at all: $\Pr_o(Z|\mathbf{h}^*_{Z=0}) = P^*_Z(Z) = \mathcal{U}(Z)$.[4]

We can use any distributional distance metric $\delta(\cdot, \cdot)$ bounded by [0, 1] to determine how far the observed distribution $\hat{P}_Z = \Pr_o(Z|\hat{\mathbf{h}}_Z)$ is from the "goal" distribution P^*_Z. Throughout this work, we use total variation (TV) distance, which allows us to directly compare counterfactual and nullifying distributions: in both cases, $0 \leq c(\hat{\mathbf{h}}_Z) \leq 1$, where attaining 1 means the intervention had its intended effect in transforming the encoding of Z. Finally, for a given set of test embeddings $\mathbf{H} = \{\mathbf{h}^k\}_{k=1}^n$, the aggregate completeness over this test set $C(\mathbf{H}_Z)$ is the average $c(\hat{\mathbf{h}}^i_Z)$ across all $\mathbf{h}^k \in \mathbf{H}$.

For **counterfactual interventions**, we measure completeness as:

$$c(\hat{\mathbf{h}}_Z) = 1 - \delta(\hat{P}, P^*_Z) \tag{1}$$

If the intervention is perfectly complete, then $\hat{P} = P^*_Z$ and $c(\hat{\mathbf{h}}_Z) = 1$. On the other hand, if $\hat{P}$ is maximally different from the goal distribution P^*_Z (e.g., $\hat{P} = \Pr_o(Z = z|\hat{\mathbf{h}}_{Z=z'}) = 1$), then $c(\hat{\mathbf{h}}_Z) = 0$. For properties with more than two possible values, completeness is computed by averaging over each possible counterfactual value $z'_1, ..., z'_k \neq z$, yielding $c(\hat{\mathbf{h}}_Z) = \frac{1}{k}\sum_{i=1}^k \hat{c}(\mathbf{h}_{Z=z'_i})$.

For **nullifying interventions**, we measure completeness as:

$$c(\hat{\mathbf{h}}_Z) = 1 - \frac{k}{k-1} \cdot \delta(\hat{P}, P^*_Z) \tag{2}$$

[1] In this paper, we use selectivity in the sense described by Elazar et al. [10], and not other probing work such as Hewitt and Liang [13], where it instead refers to the gap in performance between probes trained to predict real properties versus nonsense properties.

[2] For simplicity, we omit the superscript l denoting the layer embeddings $\mathbf{h}^l$ from which $\mathbf{h}$ is extracted; but our framework can be applied to study interventions over embeddings from any layer.

[3] An oracle probe's prediction is subtly different from the prediction an arbitrary classifier should make in the absence of any evidence about Z: such a classifier should revert to the empirical distribution $\hat{\Pr}(Z)$.

[4] This is only expected when using nullifying interventions for *causal probing* – i.e., when intervening on a model's representation and feeding it back into the model to observe how the intervention modifies its behavior. When considering nullifying interventions for *concept removal* (a more common setting), a more appropriate "goal" distribution P^*_Z would be $\Pr(Z)$, the label distribution. See Appendix B for further discussion.

where k is the number of values Z can take. The normalizing factor is needed because P^*_Z is the uniform distribution over k values and hence $0 \leq \delta(\hat{P}, P^*_Z) \leq 1 - \frac{1}{k}$.

Selectivity If an intervention on property Z_i is *selective*, the intervention should not impact M's representation of any non-targeted property $Z_j \neq Z_i$. Thus, for both counterfactual and nullifying interventions, oracle o's prediction for any such Z_j should not change after the intervention.

To measure the selectivity of a modified representation $\hat{\mathbf{h}}_{Z_i}$ with respect to Z_j, denoted $s_j(\hat{\mathbf{h}}_{Z_i})$, we can again measure the distance between the observed distribution $\hat{P} = \Pr_o(Z_j|\hat{\mathbf{h}}_{Z_i})$ and the original (non-intervened) distribution $P = \Pr_o(Z_j|\mathbf{h})$:

$$s_j\big(\hat{\mathbf{h}}_{Z_i}\big) = 1 - \frac{1}{m} \cdot \delta\big(\hat{P}, P\big) \quad \text{where } m = \max\big(1 - \min(P), \max(P)\big) \tag{3}$$

Since $0 \leq \delta(\hat{P}, P) \leq m$, we divide by m to normalize selectivity to $0 \leq s_j(\hat{\mathbf{h}}_{Z_i}) \leq 1$. If multiple non-targeted properties $Z_{j_1}, ..., Z_{j_{\text{max}}}$ are being considered, selectivity $s(\hat{\mathbf{h}}_{Z_i})$ is computed as the average over all such properties $s_{j_m}(\hat{\mathbf{h}}_{Z_i})$. Finally, analogous to completeness, the aggregate selectivity over a set of test embeddings $\mathbf{H}_{Z_i} = \{\mathbf{h}^k\}_{k=1}^n$, denoted $S(\mathbf{H}_{Z_i})$, is the average selectivity $s(\hat{\mathbf{h}}^k_{Z_i})$ across all $\mathbf{h}^k_{Z_i} \in \mathbf{H}_{Z_i}$.

Reliability Since completeness and selectivity can be seen as a trade-off, we define the overall reliability of an intervention $R(\hat{\mathbf{H}})$ as the harmonic mean of $C(\hat{\mathbf{H}}^l)$ and $S(\hat{\mathbf{H}}^l)$. This is analogous to the F1-score, which is the harmonic mean of precision and recall: just as a degenerate classifier can achieve perfect recall and low precision by always predicting the positive class, a degenerate intervention can achieve perfect selectivity and low completeness by performing no intervention at all. Using harmonic mean to calculate reliability heavily penalizes such interventions.

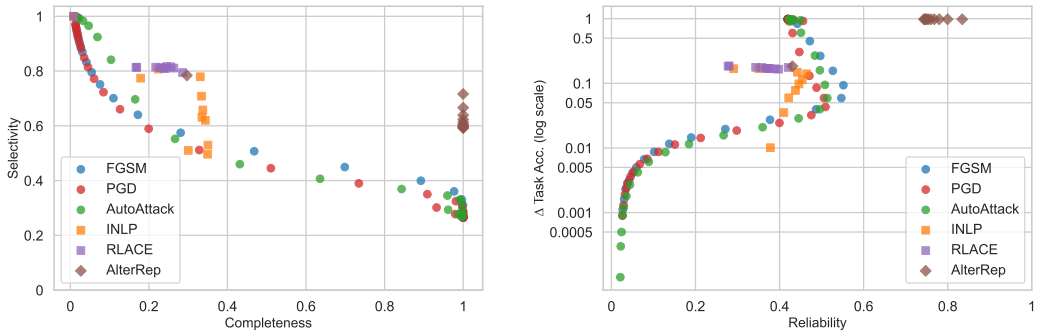
4 Experimental Setting

Language Model We test our framework by carrying out an extensive range of experiments in the context of BERT [9]. We opt for BERT because it is very well-studied in the context of language model interpretability [28], particularly in causal probing [27, 10, 25, 18, 24, 23, 7].

Task Following other causal probing works [18, 25], we select the prompting task of **subject-verb agreement**. Each data point takes the form $\langle \mathbf{x}_i, y_i \rangle$ where $\mathbf{x}_i$ is a sentence such as "the girl with the keys `[MASK]` the door," and the task of the LLM is to predict $\Pr_M(y_i|\mathbf{x}) > \Pr_M(y'_i|\mathbf{x})$ (here, that $y_i =$ "locks" rather than $y'_i =$ "lock"). The causal variable Z_c is the number of the subject (`Sg` or `Pl`), because (grammatically) this is the only variable that determines the number of the verb in English. The environmental variable Z_e is the number (`Sg` or `Pl`) of the noun immediately preceding the `[MASK]` token when that noun is not the subject (e.g., "keys" in the phrase "with the keys"). This is the simplest experimental setting (two binary properties) that allows us to study interventions using our framework; however, nothing in our methodology precludes the use of more properties, or properties with more possible values.

Dataset We use the LGD dataset [19], which consists of >1M naturalistic English sentences from Wikipedia; from this we take only sentences for which both singular and plural forms of the target verb are in BERT's vocabulary. We use 40% of the examples to train oracle probes, 40% to train interventional probes, and 20% as a test set. (More dataset details can be found in Appendix C.1.) For simplicity, in all experiments, we analyze `[MASK]` embeddings $\mathbf{h}_{\texttt{[MASK]}}$ from BERT's final layer immediately before it is fed into the prediction layer. We do this because, for earlier layers, any information about the target property Z removed or modified by interventions may be recoverable from embeddings of other tokens (see [10]), as attention blocks allow the model to pool information from the contextualized embeddings of other tokens. However, our framework is equally applicable to embeddings from any layer.

Approximating Oracle Probes Before we can measure completeness and selectivity, we must provide a suitable approximation of oracle probe o. As our goal with oracle probes is to measure the impact of causal probing interventions on the representation of the property in LLM embeddings (and not to measure how well models encode a property of interest to begin with), we follow the argument made by Pimentel et al. [22] that more expressive, higher-performing probes are a better choice. Specifically, in the results reported in the main paper, we implement $\hat{o}$ as a multi-layer perceptron (MLP) (see Appendix C.2 for further details, and Appendix D.3 for results of the same experiments



(a) **Completeness vs. selectivity** (b) **Reliability vs. ΔTask Acc**

Figure 1: **Completeness, selectivity, reliability, and ΔTask Acc** for all interventions. Each point in both plots corresponds to a different hyperparameter setting. (See Figures 5 and 6 for analogous plots with different axes for comparison.)

run with linear $\hat{o}$). Additionally, we train the oracle probes on data that is completely disjoint from that used to train interventional probes, and further ensure that Z_c and Z_e are conditionally independent in the oracle probe training set. This is crucial, as a spurious correlation between the variables could lead a probe that is trained on property Z_c to partially rely on representations of Z_e [17].

Interventions We explore two (linear) nullifying interventions: INLP [27], which iteratively trains classifiers on Z and projects embeddings into their nullspaces; and RLACE [26], which identifies a minimal-rank subspace to remove information that is linearly predictive of Z by solving a constrained minimax game. We explore one linear counterfactual method, AlterRep [25], which builds on INLP by projecting embeddings along classifiers' rowspaces, placing them on the counterfactual side of the separating hyperplanes. Finally, we study three nonlinear counterfactual methods, which are all gradient-based interventions [30, 7]: a structural MLP probe is trained on Z, then embeddings are modified to minimize the loss of the probe with respect to the target counterfactual value $Z = z'$ within an L_∞-ball of radius ϵ around the original embedding using gradient-based adversarial attacks. The attack methods (FGSM [12], PGD [20], and AutoAttack [6]) are described in Appendix C.3. After intervening on Z_c to obtain representations $\hat{\mathbf{h}}_{Z_c}$, we use approximated oracle probes $\hat{o}$ to measure completeness, selectivity, and reliability.

Impact on Model Behavior The ultimate goal of causal probing is to measure a model M's use of a property Z by comparing intervened predictions $\Pr_M(\cdot|\mathbf{x}, \mathrm{do}(Z))$ to its original predictions $\Pr_M(\cdot|\mathbf{x})$. Our framework aims to measure the reliability of the interventions themselves, a prerequisite to making claims about the underlying model. It is nonetheless important to consider how the completeness, selectivity, and reliability of a given intervention relate to its impact on model behavior. Thus, for each intervention, we also feed the intervened final-layer embeddings $\hat{\mathbf{h}}_{Z_c}$ for all test instances into BERT's prediction head, measuring task accuracy based on whether it assigns the correct verb form a higher probability, and subtract this "intervened" accuracy from the original task accuracy (98.62%) for each intervention to yield ΔTask Acc (cf. [10, 18, 7]).

5 Experimental Results

First, we note that our approximated oracle probes are able to consistently predict each property (99.4% and 88.4% accuracy for Z_c and Z_e, respectively), which is a necessary prerequisite to validate any further results.

Completeness, Selectivity, & Reliability Each intervention has a hyperparameter (ϵ for GBIs, rank r for INLP and RLACE, and α for AlterRep), where increasing its value leads to stronger interventions. Thus, each hyperparameter setting yields a different value of completeness, selectivity, and reliability for a given intervention method. Figure 1a plots selectivity against completeness for each method, showing that once each hyperparameter is raised enough, selectivity drops and reliability plateaus or decreases. See Appendix D.1 for the precise values of completeness, selectivity, and reliability at each hyperparameter value; and note that results using a linear oracle probe approximation are available in Appendix D.3, which are very similar to those using MLP oracle approximations reported in the main paper.

	$C(\hat{\mathbf{H}}_Z)$	$S(\hat{\mathbf{H}}_Z)$	$R(\hat{\mathbf{H}}_Z)$	x_{opt}
FGSM	0.8923	0.3994	0.5518	$\epsilon = 0.112$
PGD	0.7343	0.3897	0.5092	$\epsilon = 0.112$
AutoAttack	0.8433	0.3692	0.5136	$\epsilon = 0.112$
AlterRep	1.0000	0.7162	**0.8346**	$\alpha = 0.1$
INLP	0.3308	0.7792	0.4644	$r = 8$
RLACE	0.2961	0.7782	0.4290	$r = 33$

Table 1: **Intervention scores for maximum-reliability hyperparameters.** Completeness, selectivity, and reliability scores for each intervention we consider for the hyperparameter x_{opt} that maximizes the reliability of each respective method. Counterfactual methods are grouped above the double line, with nullifying methods below it.

Table 1 shows these metrics for each method at the hyperparameter that yields the highest reliability. AlterRep (the only intervention we consider that is both linear and counterfactual) achieves the highest reliability score, with perfect completeness and reasonably high selectivity. The nonlinear counterfactual methods (FGSM, PGD, and AutoAttack) all perform very similarly to each other, with high completeness and low selectivity yielding moderate reliability. The linear nullifying methods INLP and RLACE reverse this trend, with high selectivity but low completeness yielding the lowest reliability scores.

Task Accuracy Figure 1b shows ΔTask Acc as a function of the reliability for each intervention and hyperparameter setting. For most methods and hyperparameter values, ΔTask Acc increases alongside intervention reliability (largely due to changes in completeness; see Appendix D.2). There are some exceptions: first, AlterRep shows ΔTask Acc ≈ 1 for most hyperparameter values, consistent with its being the most reliable method. Second, the points at which the GBIs (FGSM, PGD, and AutoAttack) achieve the highest ΔTask Acc are *not* at their highest reliability values, resulting in a backward curve visible at the top of Figure 1b, corresponding to hyperparameter ϵ being raised past the point of maximum reliability where there is near-perfect completeness but much lower selectivity (see Figure 7). Finally, RLACE shows a similar impact on task accuracy even for different completeness and reliability scores due to its "noisy" equilibrium in reliability and completeness for high rank r (see Figure 9); whereas, similar to GBIs, INLP shows an increasing impact on task accuracy for more reliable hyperparameter settings.

6 Discussion

Tradeoff: Completeness vs. Selectivity Figure 1a shows that counterfactual methods can achieve near perfect completeness for both linear (AlterRep) and nonlinear (GBI) methods; but AlterRep has much higher selectivity. These findings support the argument made by Hewitt and Liang [13] that linear probes should be preferred because their limited expressivity prevents them from memorizing spurious associations in probe training data, leading to more selective interventions. Another possibility is that linear methods can only achieve such high completeness in these experiments because we are considering embeddings from the final layer, which have been argued to exhibit a greater degree of linearity [2].

However, despite the impressive results of AlterRep, its reliability score still peaks at 0.8346 for the optimal value of hyperparameter α, indicating that there is still room for improvement. GBIs leave even greater room for improvement: while they can be precisely calibrated to manage the completeness/selectivity tradeoff by modulating ϵ and can obtain full completeness for large ϵ values, they never achieve an overall reliability above 0.5518, meaning that improving selectivity for nonlinear interventions is clearly still an open research question. Finally, in contrast to counterfactual methods, nullifying interventions never achieve high completeness, but tend to have much higher selectivity (at least relative to GBIs). This is likely because both INLP and RLACE are explicitly optimized to minimize collateral damage [27, 26], whereas GBIs are not [30, 7].

Counterfactual vs. Nullifying Interventions All counterfactual methods (linear and nonlinear) achieve substantially higher completeness and reliability than nullifying methods (INLP and RLACE). Our evaluation framework is equally applicable to both types of interventions: it uses the same oracle probes and dataset, and is not biased in favor of one methodology or the other. Yet, consider INLP and AlterRep: in our implementation, AlterRep uses the *same* classifiers as INLP, simply carrying

out a counterfactual rather than nullifying transformation; but AlterRep has peak reliability score 0.8346 and completeness score of 1.0, while INLP has 0.4644 and 0.3503, respectively. Why should the same set of classifiers using a similar method yield such a different outcome?

One possible explanation is a general critique of causal probing: interventions (both counterfactual and nullifying) operate at the level of latent properties, meaning that they can produce embeddings that do not correspond to any specific input [11, 1]. For instance, consider the subject-verb agreement task. For input $x =$ "The girl with the keys `[MASK]` the door", what sentence $x_{Z_c=0}$ would correspond to the nullifying intervention $\mathrm{do}(Z_c = 0)$? In English, there is no corresponding noun where grammatical number is "nullified". For counterfactual interventions $\mathrm{do}(Z_c = z')$, however, there is a natural interpretation of the intervened sentence $x_{Z_c=z'}$: "girl" would be swapped out for "girls".

Given that there are no "nullified inputs" $x_{Z=0}$ that can be used to produce ground-truth "nullified embeddings" $\mathbf{h}_{Z=0}$ to train oracle probes, one natural interpretation might be that prediction over nullified embeddings is a form of "oracle probe distribution shift", which would explain the poor performance of nullifying interventions. That is, since oracle probes are never directly trained on intervened embeddings, the fact that they "transfer" better to counterfactual embeddings than to nullified embeddings is the direct result of the lack of correspondence of nullifying interventions to any particular "ground truth" (either in the input or embedding spaces) compared with counterfactual interventions. We argue that such low completeness and reliability scores for nullifying interventions are likely due to this inherent problem with the concept of embedding nullification, rather than any particular intervention methodology.

Reliability and Task Accuracy Figures 1b and 6 show a clear trend: more reliable and complete interventions and hyperparameter values show a greater impact on task performance. In particular, the most reliable intervention (AlterRep) consistently shows the greatest ΔTask Acc, and the least reliable methods (i.e., the nullifying interventions INLP and RLACE) show the least clear trend. The GBIs, which are more reliable than INLP and RLACE but less reliable than AlterRep, can damage task accuracy as much as AlterRep, but only after raising ϵ beyond its maximum-reliability setting.

This is an intuitive result: in the case where BERT does indeed perform the subject-verb agreement task by leveraging its representation of Z_c, then more reliable interventions would have a greater effect on the model's task performance. We do not claim that this is necessarily the case – e.g., results might look different if we sample from a different test distribution or intervene in earlier layers – rather, we take the clear relationship between between intervention reliability and ΔTask Acc to be a strong indicator that more reliable methods indeed yield stronger and more consistent results. This finding reinforces the utility of our framework in evaluating causal probing interventions as tools for studying models' use of latent representations.

7 Conclusion

In this work, we proposed a general empirical evaluation framework for causal probing, defining the reliability of interventions in terms of completeness, selectivity, and reliability. Our framework makes it possible to directly compare different kinds of interventions, such as linear vs. nonlinear or nullifying vs. counterfactual methods. We applied our framework to study leading causal probing techniques, finding that they all exhibit a tradeoff between completeness and selectivity. Counterfactual interventions tend to be more complete and reliable, and linear methods are generally more selective, with a linear counterfactual intervention showing the most favorable tradeoff. The large difference in reliability between a counterfactual and nullifying intervention that are otherwise almost identical suggests that the underlying differences between these intervention types, rather than differences between specific methods of each type, may explain why counterfactual interventions are more complete and reliable than nullifying methods. Finally, we observed that the most reliable interventions also had the most consistent impact on the behavior of the underlying LLM, indicating that reliability is an effective tool for comparing and evaluating intervention methods.

Acknowledgments

This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. These resources are made available through HAL [16]. Adam Davies is supported in part by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award #2229612 (National AI Institute for Inclusive Intelligent Technologies for Education). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or the U.S. Department of Education.

References

[1] Eldar D Abraham et al. "Cebab: Estimating the causal effects of real-world concepts on nlp model behavior". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17582–17596.

[2] Guillaume Alain and Yoshua Bengio. *Understanding intermediate layers using linear classifier probes*. 2017. URL: https://openreview.net/forum?id=ryF7rTqgl.

[3] Yonatan Belinkov. "Probing Classifiers: Promises, Shortcomings, and Advances". In: *Computational Linguistics* 48.1 (Mar. 2022), pp. 207–219. DOI: 10.1162/coli_a_00422. URL: https://aclanthology.org/2022.cl-1.7.

[4] Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. "Interpretability and Analysis in Neural NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Online: Association for Computational Linguistics, July 2020, pp. 1–5. DOI: 10.18653/v1/2020.acl-tutorials.1. URL: https://aclanthology.org/2020.acl-tutorials.1.

[5] Arthur Conmy et al. "Towards automated circuit discovery for mechanistic interpretability". In: *arXiv preprint arXiv:2304.14997* (2023).

[6] Francesco Croce and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks". In: *International conference on machine learning*. PMLR. 2020, pp. 2206–2216.

[7] Adam Davies, Jize Jiang, and ChengXiang Zhai. "Competence-Based Analysis of Language Models". In: *arXiv preprint arXiv:2303.00333* (2023).

[8] Adam Davies and Ashkan Khakzar. "The Cognitive Revolution in Interpretability: From Explaining Behavior to Interpreting Representations and Algorithms". In: *arXiv preprint arXiv:2408.05859* (2024). URL: https://arxiv.org/abs/2408.05859.

[9] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[10] Yanai Elazar et al. "Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals". In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 160–175. DOI: 10.1162/tacl_a_00359. URL: https://aclanthology.org/2021.tacl-1.10.

[11] Atticus Geiger, Kyle Richardson, and Christopher Potts. "Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation". In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Afra Alishahi et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 163–173. DOI: 10.18653/v1/2020.blackboxnlp-1.16. URL: https://aclanthology.org/2020.blackboxnlp-1.16.

[12] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6572.

[13] John Hewitt and Percy Liang. "Designing and Interpreting Probes with Control Tasks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2733–2743. DOI: 10.18653/v1/D19-1275. URL: https://aclanthology.org/D19-1275.

[14] John Hewitt and Christopher D Manning. "A structural probe for finding syntax in word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4129–4138.

[15] Jing Huang et al. "RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations". In: *arXiv preprint arXiv:2402.17700* (2024).

[16] Volodymyr Kindratenko et al. "Hal: Computer system for scalable deep learning". In: *Practice and experience in advanced research computing*. 2020, pp. 41–48.

[17] Abhinav Kumar, Chenhao Tan, and Amit Sharma. "Probing classifiers are unreliable for concept removal and detection". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17994–18008.

[18] Karim Lasri et al. "Probing for the Usage of Grammatical Number". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8818–8831. DOI: 10.18653/v1/2022.acl-long.603. URL: https://aclanthology.org/2022.acl-long.603.

[19] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies". en. In: *Transactions of the Association for Computational Linguistics* 4 (Dec. 2016), pp. 521–535. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00115. URL: https://direct.mit.edu/tacl/article/43378 (visited on 09/26/2023).

[20] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[21] Rowan Hall Maudslay et al. "A Tale of a Probe and a Parser". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 7389–7395. DOI: 10.18653/v1/2020.acl-main.659. URL: https://aclanthology.org/2020.acl-main.659.

[22] Tiago Pimentel et al. "Information-Theoretic Probing for Linguistic Structure". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4609–4622. DOI: 10.18653/v1/2020.acl-main.420. URL: https://aclanthology.org/2020.acl-main.420.

[23] Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. "Log-linear Guardedness and its Implications". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9413–9431. DOI: 10.18653/v1/2023.acl-long.523. URL: https://aclanthology.org/2023.acl-long.523.

[24] Shauli Ravfogel et al. "Adversarial Concept Erasure in Kernel Space". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6034–6055. URL: https://aclanthology.org/2022.emnlp-main.405.

[25] Shauli Ravfogel et al. "Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Ed. by Arianna Bisazza and Omri Abend. Online: Association for Computational Linguistics, Nov. 2021, pp. 194–209. DOI: 10.18653/v1/2021.conll-1.15. URL: https://aclanthology.org/2021.conll-1.15.

[26] Shauli Ravfogel et al. "Linear Adversarial Concept Erasure". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 18400–18421. URL: https://proceedings.mlr.press/v162/ravfogel22a.html.

[27] Shauli Ravfogel et al. "Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7237–7256. DOI: 10.18653/v1/2020.acl-main.647. URL: https://aclanthology.org/2020.acl-main.647.

[28] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. "A primer in BERTology: What we know about how BERT works". In: *Transactions of the Association for Computational Linguistics* 8 (2021), pp. 842–866.

[29] Shun Shao, Yftah Ziser, and Shay B. Cohen. *Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information*. 2022. DOI: 10.48550/ARXIV.2203.07893. URL: https://arxiv.org/abs/2203.07893.

[30] Mycal Tucker, Peng Qian, and Roger Levy. "What if This Modified That? Syntactic Interventions with Counterfactual Embeddings". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 862–875. DOI: 10.18653/v1/2021.findings-acl.76. URL: https://aclanthology.org/2021.findings-acl.76.

[31] Kevin Ro Wang et al. "Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=NpsVSN6o4ul.

[32] Andy Zou et al. "Representation engineering: A top-down approach to ai transparency". In: *arXiv preprint arXiv:2310.01405* (2023).

A Limitations

Experimental Task and LLM In this work, we focus primarily on developing an empirical evaluation framework to evaluate causal probing reliability, and deploy this framework in the context of a relatively simple model (BERT) and task (subject-verb agreement). As we discuss in Section 4, we intentionally select a more straightforward, well-studied experimental setting in order to focus on our framework and the distinctions it reveals between several types of causal probing interventions. That is, despite the proliferation of much larger and more powerful LLMs than BERT, there is still no commonly agreed-upon method for interpreting BERT's learned representations and explaining its behavior, even for simple zero-shot prompting tasks on which it achieves high accuracy. As our focus is to rigorously evaluate existing causal probing methodologies (many of which have been designed specifically with masked language models like BERT in mind), we believe it is more useful to evaluate causal probing methods in the simpler and better-understood context of BERT than it would be to scale our empirical analysis to larger models.

However, it is important to note that our evaluation framework makes no assumptions regarding LLM scale or architecture, nor the complexity of causal task structures. Thus, now that we have demonstrated the effectiveness of our framework in discovering new insights regarding leading intervention methods and providing the first hard evidence to empirically inform long-standing debates regarding causal probing, such as the conceptual issues with nullifying interventions [1, 17], we hope that our framework will be deployed and improved upon by future work to study larger and more complex tasks and models, such as autoregressive decoder-only (GPT-style) models and tasks with many more latent properties of interest.

Multiple Layers Another potential limitation of our work is that our experiments only explore interventions in the context of a single layer. As we explain in Section 4, we only examine the final layer in order to prevent information from non-intervened embeddings to be recovered by subsequent attention layers (as observed by [10]). However, our framework makes no assumptions about the specific layer to analyze – indeed, it is even possible to study the completeness, selectivity, and reliability of interventions performed in earlier layers l with respect to their impact on oracle probes in downstream layers $l' > l$. We recommend such study as an interesting direction for future work, particularly in developing approaches for understanding how information is distributed across different contextualized token embeddings and accessed by downstream attention heads (e.g., as studied in circuit discovery; [31, 5]).

Oracle Probe Approximation Finally, in our main paper, we report only results obtained using MLP oracle probes (a decision which we justify at length in Section 4). However, we also performed experiments with linear oracle probes, which we report in Appendix D.3. In general, these results to be similar: the ordering of methods and general trends remain the same, and the main differences are that INLP and RLACE have slightly better completeness and lower selectivity (yielding a marginally higher reliability score) and AlterRep also has lower selectivity (with a similar overall trend). This is not surprising, as linear oracle probes are expected to be more vulnerable to linear interventions than MLPs. Given that our goal in approximating oracle probes is to find the strongest possible probe that

is best able to recognize the model's representation, we believe that the results from the MLP oracle probes are more accurate, which is why we focus on them in the main paper.

B Framework Details

Completeness of Nullifying Interventions In Equation (2), we define the "goal" distribution P^*_Z of a nullifying causal probing intervention as being the uniform distribution – i.e., for a perfect nullifying intervention, $P^*_Z = \Pr_o(Z|\mathbf{h}^*_{Z=0}) = \mathcal{U}(Z)$. However, this is only true in the case of *causal probing*, not *concept removal*, which is a more common use case for nullifying interventions (see Section 2). That is, in the case of causal probing, the goal of an intervention is to intervene on a model's representation during its forward pass, feeding the intervened embedding back into the model and observing the change in the model's behavior (as described in Section 2). Recall that the purpose of an oracle probe o is to decode model M's representation of a given property Z, not to predict its ground truth value – that is, even if M encodes the incorrect value of $Z = z'$ rather than $Z = z$ for a given input, the oracle probe should still decode the incorrect value $Z = z'$. Indeed, this is precisely the principle behind using oracle probes in the case of counterfactual interventions that change the representation of $Z = z$ to counterfactual value $Z = z'$, where oracle probes are used to validate the extent to which the representation has actually been changed to encode this counterfactual value, and the ideal counterfactual intervention yields $\Pr_o(Z = z'|\mathbf{h}^*_{Z=z'}) = P^*(Z = z') = 1$. However, in the case of nullifying interventions $\mathrm{do}(Z = 0)$, an intervened embedding $\mathbf{h}^*_{Z=0}$ would ideally remove all information encoding M's representation of the value taken by Z, meaning that the M would not encode any value $Z = z_1$ as being more probable than $Z = z_2$ (as any information that is predictive of the value taken by Z should have been removed). In this case, the oracle probe o would predict an equal probability $\Pr_o(Z = z_i|\mathbf{h}^*_{Z=0})$ for any given value z_i that may be taken by Z_i – i.e., $\Pr_o(Z|\mathbf{h}^*_{Z=0}) = P^*_Z = \mathcal{U}(Z)$.

However, this is not the case in the context of information removal, where the goal of an intervention $\mathrm{do}(Z = 0)$ is to remove all information that is predictive of Z from embedding representations $\mathbf{h}^*_{Z=0}$ such that no probe g can be trained to predict $\Pr_g(Z|\mathbf{h}^*_{Z=0})$ any better than predicting $\Pr(Z)$ – i.e., ignoring the embedding entirely and simply mapping every input to the label distribution $\Pr(Z)$ [23]. In this case, the probe g is trained on intervened embeddings $\hat{\mathbf{h}}_{Z=0}$, in which case it can learn to map every such embedding to the label distribution $\Pr(Z)$, which yields superior performance relative to predicting the uniform distribution $\mathcal{U}(Z)$ in any case where the label distribution $\Pr(Z)$ is not perfectly uniform, as such a g would have an expected accuracy equal to the proportion of test instances with the most common label $Z = z_{\text{argmax}}$ (which would be greater than the accuracy $\frac{1}{k}$ expected by defaulting to $\mathcal{U}(Z)$).

The key technical distinction between these two use cases of nullifying interventions is whether or not probes or models are trained or fine-tuned in the context of interventions. In the case of causal probing, they are not – the (frozen) model M has no opportunity to recover the original value of $Z = z$ following a nullifying intervention $\mathrm{do}(Z = 0)$, and this should be reflected by oracle probes. This is natural, given that the purpose of causal probing is to interpret the properties used by M in making a given prediction, not to test whether M can be trained to recover properties removed by interventions; and this is reflected by oracle probes o, which are never trained on intervened embeddings. In contrast, for concept removal, probes (or models) are trained on intervened embeddings, and may learn to recover properties removed by interventions, meaning that – even in the worst case where all information has been removed – it would at least be possible to learn to reproduce the label distribution $\Pr(Z)$; but there is no reason to expect a model M or oracle probe o to do so, given that they have never been trained on intervened embeddings. Thus, while we define the "goal" distribution $P^*_Z = \mathcal{U}(Z)$ for measuring the completeness of nullifying interventions as being $\mathcal{U}(Z)$ rather than $\Pr(Z)$, this distribution would instead be $P^*_Z = \Pr(Z)$ in the case of concept removal.

C Experimental Details

C.1 Dataset

We use syntax annotations to extract values for the environmental variable Z_e from the LGD dataset [19]: if the part-of-speech of the word immediately preceding the `[MASK]` token is a noun, and it is the object of a preposition (i.e., not the subject), then its number defines Z_e. About 83% of the sentences do not have a prepositional object preceding `[MASK]`, and so are only relevant for causal interventions.

	$Z_e = \emptyset$	$Z_e = \texttt{Sg}$	$Z_e = \texttt{Pl}$	Total
$Z_c = \texttt{Sg}$	176K	31K	5K	213K
$Z_c = \texttt{Pl}$	78K	10K	4K	92K
Total	254K	41K	9K	305K

Table 2: **Contingency Table on Test Set**. Distribution of data across combinations of causal and environmental variables. $Z_e = \emptyset$ denotes instances which have no prepositional phrase attached to the subject (and thus, contain no environmental variable). Note that the label distributions are unbalanced: $\Pr(Z_c = \texttt{Sg}) = 69.8\%$ and $\Pr(Z_e = \texttt{Sg} | E \neq \emptyset) = 81.5\%$.

The contingency table for values of Z_c and Z_e in the test set are in Table 2.

C.2 MLP Probes

We do a hyperparameter sweep with grid search for the MLP probes (both oracle and interventional) that we train. The hyperparameters we consider are:

- Num. layers: [1, 2, 3]
- Layer size: [64, 256, 512, 1024]
- Learning rate: [0.0001, 0.001, 0.01]

Since the MLPs are performing classification, they are trained with standard cross-entropy loss. The probes are trained for 8 epochs, and the best probe is selected based on validation accuracy.

C.3 Interventions

Gradient Based Interventions: For all gradient-based intervention methods [7], we define the maximum perturbation magnitude of each intervention as ϵ (i.e., $||\hat{\mathbf{h}}_Z - \mathbf{h}||_\infty \leq \epsilon$), and experiment over a range of ϵ values between 0.005 to 5.0 – specifically, $\epsilon \in [0.005, 0.006, 0.007, 0.009, 0.011, 0.013, 0.016, 0.019, 0.024, 0.029, 0.035, 0.042, 0.051, 0.062, 0.076, 0.092, 0.112, 0.136, 0.165, 0.2, 0.286, 0.409, 0.585, 0.836, 1.196, 1.71, 2.445, 3.497, 5.0]$. (These are the points along the x-axis for the results visualized in Figures 1 and 4. Figures 2 and 10.) We consider the following gradient attack methods for GBIs:

1. **FGSM** We implement Fast Gradient Sign Method (FGSM; [12]) interventions as:
$$h' = h + \epsilon \cdot \text{sgn}\left(\nabla_h \mathcal{L}\left(f_{\text{cls}}, x, y\right)\right)$$
2. **PGD** We implement Projected Gradient Descent (PGD; [20]) interventions as $h' = h^T$ where
$$h_{t+1} = \Pi_{\mathcal{N}(h)}\left(h_t + \alpha \cdot \text{sgn}\left(\nabla_h \mathcal{L}(f_{\text{cls}}, x, y)\right)\right)$$
for iterations $t = 0, 1, \ldots, T$, projection operator Π, and L_∞-neighborhood $\mathcal{N}(h) = \{h' : \|h - h'\| \leq \epsilon\}$. For PGD, we use 2 additional hyperparameters: iterations T and step size α, while fixing $T = 40$, as suggested by [7].
3. **AutoAttack** AutoAttack [6] is an ensemble of adversarial attacks that includes FAB, Square, and APGD attacks. Auto-PGD (APGD) is a variant of PGD that automatically adjusts the step size to ensure effective convergence. The parameters used were set as norm $= L_\infty$ and for Square attack, the n_queries=5000.

Nullifying Interventions: For nullifying interventions, we project embeddings into the nullspaces of classifiers. Here, the the rank r corresponds to the dimensionality of the subspace identified and erased by the intervention, meaning that the number of dimensions removed is equal to the rank.[5] We experiment over the range of values $r \in [0, 1, ..., 40]$. (These are the points along the x-axis for the results visualized in Figures 4 and 12.) We consider the following nullifying interventions:

1. **INLP** We implement Iterative Nullspace Projection (INLP; [27]) as follows: we train a series of classifiers $w_1, ..., w_n$, where in each iteration, embeddings are projected into the nullspace of the preceding classifiers $P_N(w_0) \cap \cdots \cap P_N(w_n)$. We then apply the combined projection matrix to calculate the final projection where $P := P_N(w_1) \cap \cdots \cap N(w_i)$, X is the full set of embeddings, and $X_{\text{projected}} \leftarrow P(X)$.

[5]This is only true for binary properties Z – for variables that can take n values with $n > 2$, the number of dimensions removed is $n \cdot r$.

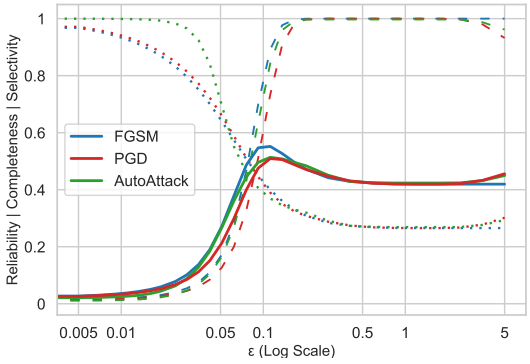


Figure 2: **Counterfactual GBIs.** Reliability (solid), completeness (dashed), & selectivity (dotted) for FGSM, PGD, and AutoAttack targeted at Z_c.

2. **RLACE** We implement Relaxed Linear Adversarial Concept Erasure (R-LACE; [26]) which defines a linear minimax game to adversarially identify and remove a linear bias subspace. In this approach, $\mathcal{P}_k$ is defined as the set of all $D \times D$ orthogonal projection matrices that neutralize a rank r subspace:

$$P \in \mathcal{P}_k \leftrightarrow P = I_D - W^\top W$$

The minimax equation is then solved to obtain the projection matrix P which is used to calculate the final intervened embedding $X_{\text{projected}}$, similar to INLP

$$\min_{\theta \in \Theta} \max_{P \in \mathcal{P}_k} \sum_{n=1}^{N} \ell\left(y_n, g^{-1}\left(\theta^\top P x_n\right)\right)$$

Hyperparameters for P and θ were a learning rate of 0.005 and weight decay of 1e-5.

AlterRep We implement AlterRep [25] by first running INLP, saving all classifiers, and using these to compute rowspace projections that push all embeddings to the positive $Z = \texttt{Pl}$ or negative $Z = \texttt{Sg}$ side of the separating hyperplane for all classifiers. That is, we compute

$$\hat{\mathbf{h}}^l_{Z=\texttt{Sg}} = P_N(\mathbf{h}) + \alpha \sum_{w \in \mathbf{W}} (-1)^{SIGN(w \cdot \mathbf{h})}(w \cdot \mathbf{h})\mathbf{h}$$

$$\hat{\mathbf{h}}^l_{Z=\texttt{Pl}} = P_N(\mathbf{h}) + \alpha \sum_{w \in \mathbf{W}} (-1)^{1-SIGN(w \cdot \mathbf{h})}(w \cdot \mathbf{h})\mathbf{h}$$

where P_N is the nullspace projection from INLP.

D Supplemental Results

D.1 Completeness, Reliability, and Selectivity

In Figure 2, Figure 3, and Figure 4, we observe that increasing the degree of control that interventions have over the representation of the target property by increasing the intervention hyperparameter associated with a given intervention type (i.e., ϵ, α, or rank) generally leads to both improved completeness and decreased selectivity.

In Figure 5, the relationship between reliability and completeness is shown analogously to Figure 1a in the main paper, but this time with reliability on the y-axis instead of selectivity. In Figure 6, the relationship between completeness and ΔTask Acc is shown analogously to Figure 1b in the main paper, but this time with completeness on the x-axis instead of reliability.

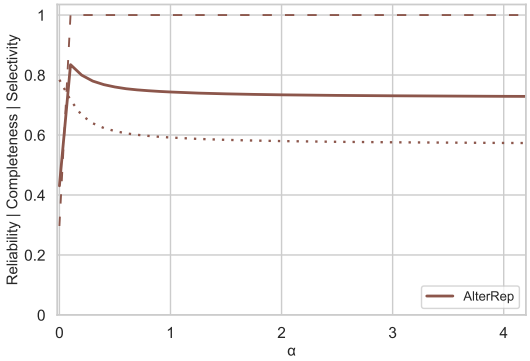


Figure 3: **AlterRep.** Reliability (solid), completeness (dashed), & selectivity (dotted) for AlterRep targeted at Z_c.

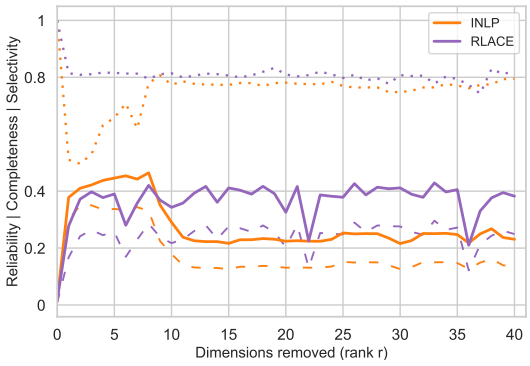


Figure 4: **Nullifying linear methods.** Reliability (solid), completeness (dashed), & selectivity (dotted) for INLP and RLACE targeted at Z_c.

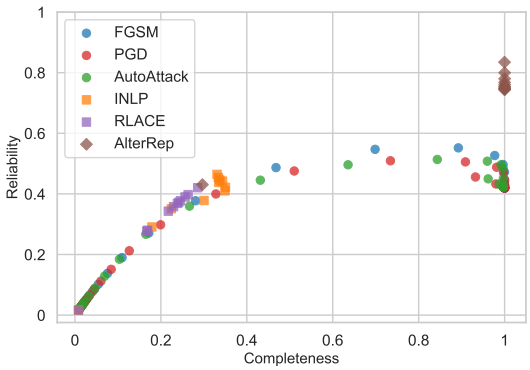


Figure 5: **Completeness vs. Reliability**

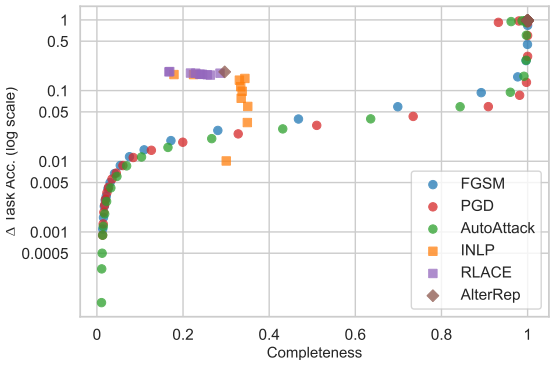


Figure 6: **Completeness vs. ΔTask Acc**

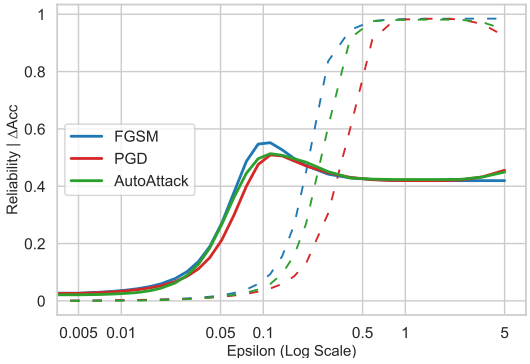


Figure 7: **Task accuracy for counterfactual GBIs.** Reliability (solid) and task accuracy (dashed) for FGSM, PGD, and AutoAttack targeted at Z_c.

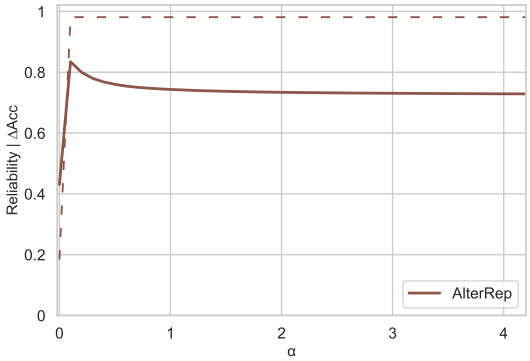


Figure 8: **Task accuracy for AlterRep.** Reliability (solid) and task accuracy (dashed) for AlterRep targeted at Z_c.

D.2 Change in Task Accuracy by Intervention Hyperparameter

Figure 7, Figure 8, and Figure 9 show reliability and task accuracy at various hyperparameter values for each method. Generally, increasing the hyperparameter values (and hence "amount of damage") results in more severe change in BERT's task accuracy.

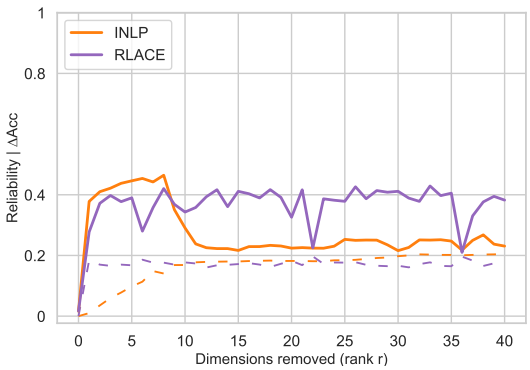


Figure 9: **Task accuracy for nullifying linear methods.** Reliability (solid) and task accuracy (dashed) for INLP and RLACE targeted at Z_c.

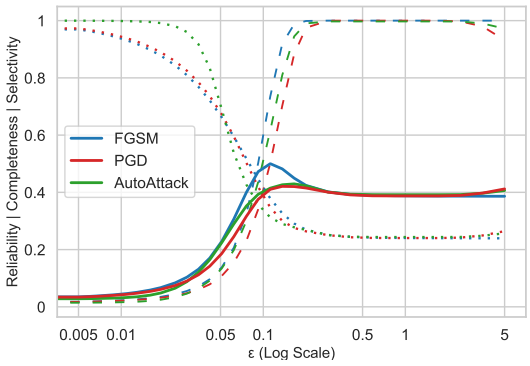


Figure 10: **Linear oracle probe on counterfactual GBIs.** Reliability (solid), completeness (dashed), & selectivity (dotted) for FGSM, PGD, and AutoAttack targeted at Z_c, using *linear* oracle probes for evaluation.

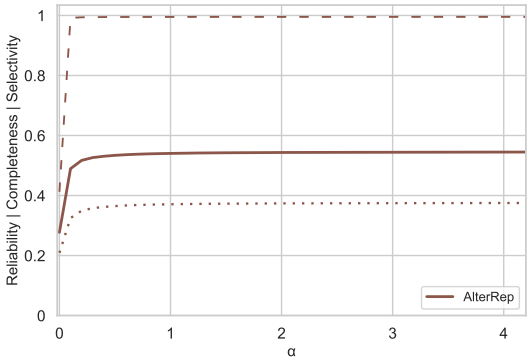


Figure 11: **Linear oracle probe on AlterRep.** Reliability (solid), completeness (dashed), & selectivity (dotted) for AlterRep targeted at Z_c, using *linear* oracle probes for evaluation.

D.3 Linear Oracle Probes

In this section, we present plots depicting reliability, completeness, and selectivity where the oracle probe has a *linear* architecure (not MLP) for each intervention. These linear probes are trained with cross-entropy loss, and a grid search was performed over learning rates in [0.0001, 0.001, 0.01, 0.1] to find the hyperparameter with the lowest validation-set accuracy. The plots are shown in Figures 10, 12, and 11.

For counterfactual GBIs, the linear oracles present very similar results to those computed with MLP oracles, shown in Figure 2. AlterRep also shows similar results to those presented in 3, except that selectivity is lower: this is expected, as linear oracle probes should be expected to be less resilient to linear interventions than MLP oracle probes. INLP and RLACE show a lower selectivity and higher completeness when evaluated with linear oracle probes versus MLP oracles, for the same reason.

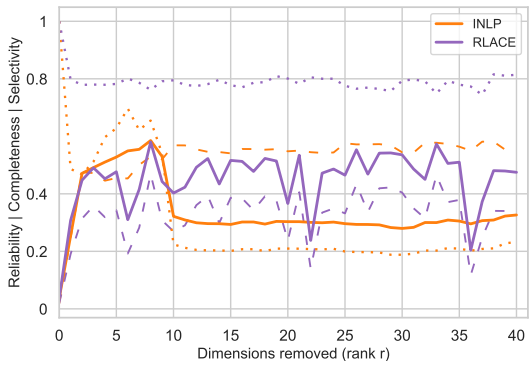


Figure 12: **Linear oracle probe on nullifying linear methods.** Reliability (solid), completeness (dashed), & selectivity (dotted) for INLP and RLACE targeted at Z_c, using *linear* oracle probes for evaluation.